DOCUMENT RESUME

ED 050 169 TM 000 565

AUTHOR Tuckman, Howard P.
TITLE The Use of Predictive Models in Forecasting Student Choice.
INSTITUTION Florida State Univ., Tallahassee. Inst. for Social Research.
PUB DATE Feb 71
NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS College Choice, Cross Cultural Studies, Cultural Factors, *Educational Benefits, Ethnic Groups, Higher Education, High Schools, *Income, *Models, *Multiple Regression Analysis, Post Secondary Education, Predictive Measurement, *Predictor Variables, Probability Theory, Seniors, Socioeconomic Status, Student Costs, Student Research

ABSTRACT

This paper uses ordinary least squares regression to obtain probabilities for the post-graduation choices of high school seniors, and it presents an illustration of the use of these probabilities in calculating future income. Problems raised by the use of the least squares regression are discussed. The benefits of higher education and ways in which they may be used as predictors are considered. The estimates presented are based upon data collected by a questionnaire administered to 2453 public high school seniors in Dade County, Florida. (GS)

ED050169

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

The Use of Predictive Models in Forecasting Student Choice

Howard P. Tuckman*

To date, most research on post-high school plans of seniors has concentrated on identifying the determinants of college choice or has simply compared the characteristics of college and non-college students. But our understanding of the selection process increases if several post-high school alternatives are examined. This paper uses ordinary least squares regression to obtain probabilities for the post-high school choices of high school seniors, and it presents an illustration of the use of these probabilities in calculating future incomes. In the process, we examine several problems raised by the use of the least squares regression.

Herbert Simon suggests that "the viewpoint is becoming more and more prevalent that the appropriate scientific model of the world is not a deterministic model but a probabilistic one."[1] In analyzing post-high school choices, the challenge becomes one of finding a set of variables which provide reasonably accurate predictions and of clearly identifying the post-high school alternatives considered by students. In the next section, we consider the benefits of higher education and several ways that these benefits can be used as predictors. We then use an adjusted

TM 000 565

---

*The author is an Assistant Professor in the Department of Economics and a Research Associate in the Institute for Social Research at the Florida State University. He gratefully acknowledges the research assistance of Victor Steeb, the programming assistance of Adele Spielberger, and the helpful comments of Warren Mazek and John Chang.

ordinary least squares (OLS) technique to estimate post-high school choices for 2,453 Dade County seniors. The conditional probabilities obtained are then utilized to predict the future incomes of high school seniors with select characteristics.

Benefits of Higher Education as a Predictor of Choice

Although several economists have examined the benefits of schooling, to my knowledge, few have attempted to demonstrate empirically that these benefits have an effect on student choice. The rapidly expanding literature on human capital, while useful in linking investment in physical and human capital, has not as yet succeeded in establishing that rates of return from schooling affect a student's post-high school plans. Moreover, no methods have been devised to separate the effects of current and future benefits on student's choice of an educational path.

In this section we consider three alternative methods for incorporating perceived benefits in a predictive model. These approaches are not exhaustive and several others might have been discussed. My major purpose here is to establish the importance of these benefits in determining choice.

Net Benefits as a Predictor of Choice

The net benefits approach begins with the assumption that students select that educational path that offers the greatest net benefits to them. We shall use the term "benefit" to refer to anything that increases the satisfaction of the student.[2] Included in our definition would be anything improving the student's future earning power, anything increasing his current satisfactions or reducing the costs that he might otherwise

have to incur. The value of the benefits a student expects to receive from further education depends upon his time preferences for current and future returns, the way he weighs the various goods and services provided by institutions of higher education and upon the motivation, intelligence and attitudes he possesses.

We can identify several benefits of education. One benefit is the financial return received from further education. Second, we might include the "financial option" return which involves the opportunity for students to obtain still further education. Goods and services received by the student while he obtains his education provide a third source of benefits. Moreover, non-market benefits and "opportunity options" (i.e., the additional employment opportunities offered to the educated) might also be identified. If a student takes these benefits into account in making post-high school plans then an increase in one or all of them should increase the probability that he continues his education.

The financial returns from higher education are well known although they may be measured several ways.[3] A careful formulation of these requires the researcher to allow for ability and motivational differences among students, and for other characteristics affecting future earnings.

Weisbrod has suggested that completing one level of schooling provides a student with the opportunity to continue on to further education.[4] The "financial option" attributes to investment in one educational path a fraction of the net financial returns from further education. This approach suggests that the benefits of completing a four-year college should include the value of the financial option to continue on to graduate school. The

value of an option to pursue further schooling depends upon the likelihood that it will be exercised and upon its expected value when exercised. Moreover, this value will be greater the further down the educational ladder the student is. While Weisbrod has shown that a monetary value can be assigned to the financial option there is no evidence to indicate that this option plays a significant role in the choice process.

A student selecting among colleges or college types considers different bundles of goods and services. Some empirical evidence of this has been compiled by Astin.[5] Colleges, and to a lesser extent, other forms of higher education produce both investment and consumption services. In deciding whether to attend a college and which college to attend, a student chooses among the bundles of goods and services available at each of several colleges and other goods provided by the market place. For example, a student interested in finding good books to read may do so through the market place by joining one of the many "great books" clubs. Thus, if all he desires is the opportunity to read, he need not go to college. Similarly, a student interested in a college education and a good social life may find this several ways. He might go to the local junior college and join a local social club, or he might choose to attend a social university away from home.

Some services are provided both by the market place and by colleges. Others may only be accessible to college attendees. If the former are influential in determining whether a high school student chooses a college career as against an alternative path, then a rise in the price of college relative to market price diminishes the likelihood of college attendance.

If market services are poor substitutes for college services, a change in their price will have little effect on high school student choice.

Consider two alternative methods for valuing the goods and services (hereafter called services) received at college. The first assumes that the value of the services received by a student must at least equal the price he pays to attend the college. This method provides no guidelines for separating investment and consumption purchases. Alternatively, if the services received by a student can be identified, they might then be valued using the price of equivalent services in the private sector. Thus, current services could presumably be included either in a single net benefits measure or they might be entered as separate parameters in a model used to predict student choice.

Higher education increases the alternative job possibilities open to the student and permits him to choose among a larger and more varied number of jobs. Several of these jobs involve non-monetary rewards. For example, the clergyman, social worker, or civil servant may receive benefits from their jobs in the form of psychic satisfaction or security. While these are difficult to quantify, they no doubt figure in some students' choice of educational path.

Non-market benefits pose valuation problems both for the researcher and for the student. Questions arise concerning the value of being away from home, of maturity before entering the labor force, of exposure to a broad range of new ideas, etc. These benefits are important insofar as they are taken into account by the decision-maker. Since students (and their parents) have limited knowledge of the future, it seems reasonable

to assume that benefits of this type will be ignored or at least heavily discounted.

The above discussion suggests the difference in the approach taken by economists and psychologists. Since economists take tastes as given the preferred variables in their models are the prices of alternative goods and services. Psychologists, however, utilize personality traits to predict choice. Both approaches could be used to examine the importance of net benefits as a predictor of choice.

Consider two alternative possibilities for operationalizing the net benefits approach. One involves systematic identification and cataloging of benefits at different colleges. For example, students might be asked to indicate the benefits they expect to receive at each of several schools. We might then test the hypothesis that a student would choose that school where the net benefits (total benefits less costs) were greatest.

Alternatively, statistical techniques such as regression or discriminant analysis might be used to predict the weights given to each set of benefits. Work by Holland and Nichols, and Holland has demonstrated that the post-high school plans of seniors are related to a variety of personality characteristics.[6] It may be possible to link the benefit selection procedure directly to these characteristics but further attention must also be given to providing a set of benefits that have meaning in the market place.

Price as a Predictor of Choice

Economic theory assumes that individuals maximize their utility subject to an income constraint. Given a set of services available in

the economy, the rational consumer chooses among them according to the satisfactions they yield. Eventually he arrives at an equilibrium point by allocating his income among the services in such a way as to insure that no other set of services leaves him better off.

We have already noted that several services are purchased by a student seeking further education. Unless the student does not receive satisfaction from these services his decision as to whether to attend a college will depend upon the price of the college and on the price of equivalent services which are available outside the college. This suggests a demand function of the form

$$(1) \quad D = f(P_c, P_1, \ldots, P_j, \ldots, P_n)$$

where the demand for a college is a function of its own price (P_c) and of the price of equivalent services elsewhere (P_1 to P_n). Since equivalent services may be supplied at other colleges some of the P_j's represent prices of other colleges. Using price theory we might then forecast that the quantity of higher education demanded will increase when the price of that education decreases or when the cost of equivalent services increases.

In a recent paper based upon data for California high school seniors, Hoenack showed that meaningful predictive equations can be estimated for colleges in the University of California (U.C.) system. Unfortunately, he was unable to extend his analysis to include private and public non-U.C. schools and thus failed to include the prices of schools outside the U.C. system in his equations. Moreover, his formulation failed to provide a method for predicting individual behavior but rather utilized the high school as basis for prediction. Hoenack's method is of some interest to

those wishing to predict demand for state colleges but it is unsatisfactory for describing choice among a broader range of higher education alternatives.

Operationalizing the price approach poses several problems. The services provided by institutions are not obvious nor is it clear which services are complements and which substitutes. While researchers can study student choice over a period of time to determine which prices affect choice the task will not be easy. Among Dade County students, for example, we found that almost 82% of the students applying to one school failed to apply to a second and 85% failed to apply to a third.[7] These findings suggest that students may first set an acceptable price and then shop for a college. Alternatively, they may suggest that students lack information about alternatives. In either event, they leave the researcher with little information as to which prices are the relevant ones to utilize in empirical studies.

If the prices of market-provided services are to be included in a predictive model they must first be identified. By assuming that the value of benefits received by a student while at college varies directly with the prices of consumer goods, one might use the consumer price index to reflect changes in market-provided goods. But while a general rise in consumer prices raises the cost of obtaining similar benefits outside the college environment the net effect of this on demand will depend upon whether college price rises less than the general price index.

Student Preferences as a Predictor of Choice

A third method of capturing benefits utilizes student responses to indicate whether college activities are desirable. If students desire the

current and/or future services attainable at colleges they will be more likely to choose a college path than if they are interested only in market-provided services. Moreover, if four-year colleges differ from junior colleges, and if junior colleges differ from vocational schools in the type of services they provide, then information on the activities desired by a student will aid in predicting his choice. While this method does not capture all of the benefits suggested in the benefits approach, it provides a first approximation to the effects we are after.

The difficulties of this approach are well known. Student responses may be dependent upon the wording of the question. Moreover, students may differ in the degree of certainty with which they hold a view. Further, in some cases it may be desirable to work with a large number of questions dealing with different aspects of higher education environments rather than relying on a single measure.

In attempting to formulate the model in the next section several alternative benefit measures were considered. Our objective was to formulate a reasonable set of predictors bearing some resemblance to services provided by higher education institutions. Several indices of current services were considered and dropped in favor of a set of dummy variables consisting of the student's expressed desire for: 1) Social activities such as clubs, Greek organizations, dating and dancing; 2) Special interest activities such as art, theater, journalism or photography; 3) Political activities including both student government and/or radical politics; 4) Future career related activities such as membership in a business honorary or opportunities to work with future employers; 5) College sports of either

a participatory or viewing nature; and 6) All other activities (such as getting to know other people, living with people of the same age, etc).

In the following sections we report the results of an attempt to use OLS regression to determine probabilities for the alternative educational choices. These findings should be viewed as preliminary since the model has not been tested for interaction effects and for prediction bias. The former problem is especially important and has been explored in a somewhat similar context elsewhere.[8]

Data Description

The estimates presented in this paper are based upon data collected from high school seniors in Dade County (Miami) Florida. A questionnaire was administered to a random sample of high school seniors intending to graduate in the spring of 1970. Completed questionnaires were obtained from 18% or 2,453 of the 13,542 high school seniors enrolled in the public school system. Initial plans called for a 20% random sample but some attrition occurred due to absenteeism and to incomplete or improperly completed questionnaires. The methods used for selecting the sample and a more detailed analysis of student responses may be found in a recent study by Grigg, et al.[9] Means and standard deviations of the variables used in the regression appear in Appendix Table 1.

The following variables were selected from among the variables shown in Appendix Table 2 using the extra sum of squares principle.[10] They explain a significant amount of the variation in at least one regression.

x_1 - a dummy variable denoting that a student is male

x_2 - a dummy variable denoting that a student is black

x_3 - a dummy variable denoting that a student is Cuban

x_4 - a dummy variable denoting that a student ranks in the top quarter of his class

x_5 - a dummy variable denoting that a student ranks in the second quarter of his class

x_6 - a dummy variable denoting that a student ranks in the third quarter of his class

x_7 - a dummy variable denoting that a student's father has some education beyond high school

x_8 - a dummy variable denoting that a student's father has a college degree

x_9 - a dummy variable denoting that father's income falls between $3,000 and $4,999

x_{10} - a dummy variable denoting that father's income falls between $5,000 and $9,999

x_{11} - a dummy variable denoting that father's income falls between $10,000 and $14,999

x_{12} - a dummy variable denoting that father's income is $15,000 or more

x_{13} - a dummy variable denoting a major interest in college social life

x_{14} - a dummy variable denoting a major interest in college special interest activities

x_{15} - a dummy variable denoting a major interest in college political activities

x_{16} - a dummy variable denoting a major interest in college future career related activities

x_{17} - a dummy variable denoting a major interest in college sports

x_{18} - a dummy variable denoting a major interest in "other" college activities

x_{19} - a dummy variable denoting that father or mother exerted a major influence on the student's choice

x_{20} - a dummy variable denoting that counselor or teacher exerted a major influence on the student's choice

x_{21} - a dummy variable denoting that a friend exerted a major influence on the student's choice

x_{22} - a dummy variable denoting that a relative or some other person exerted a major influence on the student's choice

Several prior studies have shown the importance of sex and race in determining college choice.[11] Moreover, a large body of literature supports the importance of high school grades.[12] Where researchers have laid SES measures aside, both father's income and education appear to affect college choice.[13] Although the presence of both variables might result in some multicollinearity in the model this does not seem to be a serious problem when viewed in terms of the zero-order correlations of the variables.[14]

We have utilized father's income as reported by the student. Although this measure may understate "true" income a comparison of father's income and occupation as reported by students with census estimates suggests that the means and standard deviations of the reported income variable are reasonable.[15] Both the activities variables and the variables denoting major influence on the student come from student responses and are subject to the difficulties discussed in the last section. Finally, the importance of parental and peer group influence has been documented by several researchers.[16]

Studies using aggregated data for prediction and applied to more detailed college classifications have found that college prices affect student choice.[17] Experiments with several functional forms and with

several formulations of the price variable did not produce a meaningful effect when a price variable was added to the model. Since some students not interested in colleges may have considered a college but did not indicate this fact on the questionnaire, prices could only be obtained for those planning on attending a two or four-year college. This resulted in a positive sign on the price coefficient in the regression predicting choice of a four-year college. One interpretation of this sign is that including price in the model raises R^2 in much the same way that placing the dependent variable on both sides of the equation would. Another interpretation is that the price variable acts as a proxy for college benefits. A rise in college price indicates that expected benefits increase. However, the latter hypothesis does not explain the negative sign on the price variable in the other regressions.

An important difference between this study and several others[18] is that it estimates conditional probabilities for several post-high school alternatives including:

y_1 = no further education beyond high school

y_2 = attend a business or vocational school

y_3 = attend a junior college

y_4 = attend a junior college then continue to a four-year school

y_5 = attend a four-year college

Some students were unsure of their future plans and these students were included in the sample used to estimate the regressions so as to provide a 6th alternative of uncertainty. Together, the six categories cover all alternatives available to the student.

Regression Results

For each individual in the sample and for each alternative educational choice we estimated a simple linear regression equation of the form

$$(2) \quad y = \alpha + \sum_{i=1}^{22} b_i x_i + u$$

where y and the x_i have been defined above and u is a random disturbance term.[19] Non-significant terms were then removed using addition to R^2 as a criterion for removal and the equation re-estimated. Appendix Table 3 gives zero order correlations for the variables while the regression results appear in Table 1.

The left-hand column of Table 1 lists the independent variables. Note that "regression intercept" represents α in equation (2) and that since the model is in dummy variable form, each of the b_i coefficients represents an addition to or subtraction from the intercept term. To obtain the coefficients for a particular regression equation, one reads down the appropriate column. The regression intercept for the four-year college regression is -.008. If a student is male this adds .032 to the intercept and if his father has a college education this raises the intercept by an additional .158.

[Table 1 about here]

Note that the signs on the regression coefficients and their direction of change support our a priori expectations. For example, consider the high school rank variables contained in the first regression. As a student's rank rises, the probability that he will stop his education decreases. An increase in high school rank also lowers the probability of attendance at

TABLE 1

Regression Estimates of the Effects of the Independent Variables on Various Career Choices

Variable	Educational Choice				
	No Further Education	Business or Vocational School	Junior College	Junior College and Second 2 Year College	Four Year College
Regression Intercept	.248	.185	.134	.049	-.008
Male	-.044 (.010)		-.073 (.014)	.097 (.018)	.033 (.015)
Black	-.068 (.015)		-.090 (.020)	-.077 (.025)	-.011 (.022)
Cuban	-.056 (.016)			.115 (.028)	-.084 (.024)
High School Rank					
Top 25%	-.066 (.016)		-.080 (.022)	-.027 (.029)	.316 (.024)
25-50%	-.046 (.016)		-.002 (.022)	.060 (.028)	.070 (.024)
50-75%	-.024 (.015)		.003 (.021)	.040 (.027)	.006 (.023)
Father's Education					
Some Beyond High School					.084 (.019)
College Degree	-.019 (.013)	-.056 (.015)	-.048 (.017)		.158 (.010)
Father's Income					
$3,000-$4,999	.073 (.021)	-.014 (.021)	.053 (.028)		-.037 (.030)
$5,000-$9,999	-.006 (.016)	.029 (.019)	.039 (.022)		-.039 (.023)

TABLE 1--Continued

$10,000-14,999	-.014 (.015)	.017 (.018)	.049 (.021)		-.064 (.023)
$15,000 +	-.038 (.015)	-.018 (.018)	.016 (.021)		.040 (.023)
Desired Activities					
Social	-.113 (.016)	-.153 (.019)		.172 (.028)	.153 (.024)
Special Interest	-.113 (.018)	-.133 (.022)		.183 (.032)	.127 (.027)
Political Act.	-.103 (.031)	-.150 (.037)		.118 (.055)	.248 (.046)
Future Career	-.086 (.017)	-.134 (.020)		.137 (.030)	.053 (.026)
Sports	-.092 (.014)	-.149 (.017)		.126 (.026)	.130 (.021)
Other	-.074 (.023)	-.093 (.028)		.065 (.042)	.036 (.035)
Major Influence on Student					
Father or Mother	-.056 (.015)		.054 (.021)	.098 (.027)	.015 (.022)
Counselor or Teacher	-.070 (.025)		.060 (.034)	.105 (.044)	-.038 (.036)
Friend	.046 (.020)		.109 (.028)	-.020 (.036)	-.046 (.029)
Relative or Other	-.024 (.021)		.030 (.029)	.078 (.037)	.002 (.031)
Multiple Correlation Coefficient (R^2)=	.12	.07	.05	.08	.25
F-Test =	16.2	17.3	7.9	12.9	35.9
Standard Error =	.24	.29	.33	.43	.36

a junior college while raising the probability that a student attends a four-year college. Thus, our results seem to be consistent. Similarly, a low father's income raises the probability that a student stops and lowers the probability of his going to a four-year school while a father's income in excess of $15,000 has the reverse effect.

In general the coefficients on the high school rank and desired activities variables are larger than those of any other variables in the regressions where they appear. If the b coefficients are converted to beta coefficients to provide "standardized" regression weights, the high school rank coefficients become larger while the desired activities terms are no longer as important as a student's sex.[20] Interestingly, desired activities have an important effect on the probability of choosing all educational alternatives except junior colleges. This finding seems especially interesting since the sign on the activities variable is negative in the first two regressions and positive in the second two. Perhaps junior colleges provide benefits which neither attract students interested in college benefits nor repel them.

As expected, the regression equations differ in their ability to explain variations in the dependent variable. The regression equation for four-year colleges best fits the data (R^2 = .25) while the worst fit (R^2 = .05) occurs for the regression explaining choice of junior college. While all of the R^2's obtained seem low, this is not uncommon for regression estimates obtained from disaggregated data.

Before using the regression coefficients in Table 1 to obtain conditional probabilities, we must first adjust our estimates. If no adjustment is made to the summed coefficients, then the conditional probability

obtained may fall outside the 0-1 range normally associated with probabilities. It may only take one reason to convince a student not to go to college (i.e., low grades) so that other factors discouraging attendance (i.e., such as low family income) may make the actual probability of attending college less than zero. Thus, a less than zero estimate occurs due to the additivity assumption made in equation (2).

[Table 2 about here]

To remove the worst effects of the additivity assumption, we utilize a method suggested by Orcutt.[21] The fitted values (y*) from the regression are grouped into relatively homogeneous categories and a mean and standard error of the residuals (i.e., the difference between the actual value and the fitted value y*) are computed.[22] A t-test is then used to determine whether the residual estimate of a category differs significantly from zero. The user sums his coefficients and looks up the mean and residual for this value of y*. If the mean is significant, he adds it to the y* estimate. If not, then the summed coefficients provide a correct estimate of conditional probability. Table 2 gives adjustment factors and their standard errors for each of the regressions presented above. An asterisk denotes that the t-test suggested a mean significantly different from zero.

Using the regression coefficients in Table 1 and the adjustments in Table 2, we can predict the post-high school choices of students with different characteristics. From among several alternative combinations of coefficients, four situations commonly found in the sample have been chosen. Entry I of Table 3 shows the probability that a male student in the top 25% of his class with a father earning more than $15,000, interested in sports,

TABLE 2

Adjustments of the Residual Terms

No Further Education			Business or Vocational School		
Range of Y*	Mean of Residuals	Standard Error	Range of Y*	Mean of Residuals	Standard Error
Less than-.100	.116*	.004	-.073 to .006	.031*	.002
-.100 to -.035	.068*	.003	.006 to .031	.005*	.002
-.035 to -.003	.026*	.004	.031 to .054	.002*	.001
-.003 to .027	-.003	.002	.054 to .069	-.032*	.006
.027 to .065	-.036*	.008	.069 to .159	-.029*	.006
.065 to .100	-.037	.009	.159 to .171	-.008*	.004
.100 to .133	-.062	.013	.171 to .202	.005	.004
.133 to .177	-.001	.001	.202 to .215	.043*	.011
.177 and above	.061*	.014	.215 and above	-.006	.005

Junior College			Junior College Plus Second Two Years		
Range of Y*	Mean of Residuals	Standard Error	Range of Y*	Mean of Residuals	Standard Error
-.069 to .051	.029*	.002	-.055 to .118	.037*	.001
.051 to .081	-.004*	.001	.118 to .188	-.013*	.002
.081 to .113	-.026*	.004	.189 to .228	.014*	.003
.113 to .137	-.021*	.005	.228 to .285	-.024*	.004
.137 to .158	-.020*	.005	.285 to .312	-.010*	.003
.158 to .182	-.008*	.003	.312 to .358	-.062*	.001
.182 to .218	.039*	.008	.358 to .416	.010*	.004
.218 and above	.012*	.004	.416 and above	.048*	.008

Four Year College Choice		
Range of Y*	Mean of Residuals	Standard Error
-.177 to -.008	.071*	.005
-.008 to .046	.025*	.005
.046 to .103	-.011*	.004
.103 to .168	-.046*	.011
.168 to .244	-.052*	.013
.244 to .347	-.054*	.014
.347 to .469	-.002	.003
.469 to .802	.067*	.017

and with father or mother influencing his choice, will choose each of the educational alternatives. Column 1 presents probabilities for whites, while Columns 2 and 3 give probabilities for blacks and Cubans (based upon the regression coefficients in Table 1). Note the addition of the "unsure" category to the set of educational alternatives. We have not estimated a separate probability for an "unsure" answer since the category can be obtained as one minus the sum of the other probabilities.

[Table 3 about here]

The probability that a white student will choose a junior college rises as his father's income level and his own rank in high school fall. This also holds true for Cubans and blacks, although the probability increases more for Cubans than for other groups. Moreover, the probability that a student will attend a four-year college decreases as his high school rank and his father's income fall. Blacks from $15,000+ families and in the top 25% of their class choose four-year colleges as frequently as whites. While this also holds true at lower income levels, a greater percentage of blacks are unsure of their plans and fewer plan to attend junior college as compared to whites. These results appear to be consistent with an earlier finding, using Coleman report data, that being black is not negatively associated with college attendance but rather with completion of high school.[23]

Junior colleges are favored by Cubans of almost all income levels. One explanation of this may be a desire among Cubans to live in the Miami area.[24] While Cubans are not as likely to choose a four-year college as whites or blacks, they are at least as likely to select a four-year

TABLE 3

Probability of A Male Student Continuing His Education

I. Male Student with A Father Whose Income Exceeds $15,000, in the Top 25% of His Class, with Family Influencing His Educational Choice

	Predicted Probability		
	White	Black	Cuban
No Further Education	.02	.00	.00
Business or Vocational School	.02	.02	.02
Junior College	.05	.02	.05
Junior College Plus Second 2 Years	.28	.24	.50
Four Year College	.59	.58	.43
Unsure	.04	.14	.00

II. Male Student with A Father Whose Income Falls Between $10,000-14,999, in the Top 25% of His Class with Family Influencing His Choice

No Further Education	.00	.00	.00
Business or Vocational School	.05	.05	.05
Junior College	.06	.02	.06
Junior College Plus Second 2 Years	.28	.24	.51
Four Year College	.42	.41	.28
Unsure	.19	.28	.10

III. Male Student with A Father Whose Income Falls Between $5,000-9,999, in the Second Quarter of His Class with Family Influencing His Choice

No Further Education	.06	.00	.01
Business or Vocational School	.03	.03	.03
Junior College	.13	.06	.13
Junior College Plus Second 2 Years	.49	.29	.59
Four Year College	.15	.14	.07
Unsure	.14	.48	.17

IV. Male Student with A Father Whose Income Falls Between $3,000-4,999, in the Third Quarter of His Class, Influenced by A Friend

No Further Education	.43	.38	.36
Business or Vocational School	.02	.02	.01
Junior College	.21	.12	.19
Junior College Plus Second 2 Years	.28	.23	.38
Four Year College	.06	.06	.06
Unsure	.00	.19	.00

Table Notes: We use a .00 to denote a probability of zero. Although the groupings used in the Orcutt adjustment should be based upon homogeneous observations, this is not completely possible. Thus, the possibility exists that negative probabilities may occur at group boundary points. Slightly negative probabilities are found in the "no further education" category but the differences are so small that no reordering of the groupings was undertaken.

education by utilizing the junior college plus two additional years of schooling.

While the models presented above seem to capture the factors influencing why students continue their education they add little to our knowledge of why students do not continue their education. The intercept term on the first regression indicates a 25% probability that students will not continue. This probability is unrelated to the variables in the model and suggests that it would be fruitful to learn more about the causes of non-continuation. Moreover, none of the significant variables except low father's income and friend's influence contributed positively to the probability of non-continuation while almost all contributed negatively. More work must be done to identify positive influences on non-completion.

Our results are preliminary. The probabilities calculated from the above regressions should be tested against the probabilities observed in the cells of the sample. Moreover, our models should be tested on a broader sample of students.

A rough comparison with Schoenfeldt's study of Project Talent data suggests that our estimates seem reasonable.[25] Although Schoenfeldt used an ability measure based upon 16 aptitude and ability scores and a socio-economic status (SES) measure obtained from nine different items, his results seem to be similar to ours. For example, 85% of those in his top ability and SES quartiles planned to attend a four-year college.[26] Adding our estimate of the probability of attending junior college and continuing for a second two years to the probability of attending a four-year college in entry I of Table 3 gives us an equivalent estimate of 86%. Similarly,

Schoenfeldt estimates that moving from the top ability quartile to the second quartile decreases the probability of four-year college entry by 24%. This appears to be consistent with our estimated decrease from 32% to 07%, obtained from the last column of Table 1 by moving from the top 25% rank in high school to the 25-50% rank. Our estimates for college attendance for the lower income groups appear to be higher than Schoenfeldt's but this may be due to the open door policies followed by the Miami-Dade Junior College and to our separation of ethnic groups. On balance, the estimates presented here appear to be plausible when compared to those obtained from Project Talent data.

An Application of Conditional Probabilities to the Problem of Predicting Future Incomes

Economists have increasingly recognized that the supply of human capital (the stock of skills and knowledge of an individual) has an important effect on the distribution of income. Studies of the process of human capital accumulation have identified two important effects of student background characteristics. The first involves the effect of socio-economic and other background characteristics in determining the educational level.[27] For example, low family incomes may cause students to drop out of high school. The second involves the extent to which the returns to graduates from institutions of higher education depend upon their backgrounds, abilities, and other factors.[28] For example, high ability students may earn more as a result of their education than average students. Since inequalities of income stem from both of these sources, it will be useful to describe this process in a simple model.

Let $\bar{E}_i$ denote the expected lifetime income of an individual with a set of i characteristics. Let p'_{ij} represent the estimated probability that an individual with characteristics (i) will complete a particular educational path (j) and E_{ij} represent the expected lifetime income for an individual with characteristics (i) completing path (j). Now p'_{ij} consists of two other probabilities; y'_{ij} the conditional probability of choosing an educational path as estimated in the last section, and d'_{ij} the probability that an individual will complete this path so that $p'_{ij} = y'_{ij}d'_{ij}$. For simplicity, let $d'_{ij} = 1$ for all j so that we may estimate the lifetime earnings of an individual as follows:

$$(3) \quad \bar{E}_i = \Sigma y'_{ij}E_{ij}$$

The probabilities estimated earlier thus enable us to estimate the expected future income of high school seniors with different characteristics. Utilizing the probabilities calculated for the entries in Table 3 and an estimate of expected lifetime incomes we have calculated expected lifetime incomes for the students in our sample. Table 4 gives the results of our calculations.

[Table 4 about here]

In order to estimate expected lifetime incomes it was necessary to obtain information on cross-section incomes for individuals with different educational backgrounds. A recent Census Bureau publication presents lifetime income estimates for those aged 18 in 1968.[29] These estimates assume constant 1968 dollars and allow for adjustments for productivity and the probability of survival.[30] We have chosen to use age 18 as the basis for our estimates since this corresponds to the age of most high school graduates.

TABLE 4

Expected Discounted Future Income for High School Graduates

	Whites	Cubans	Blacks	
			I	II
Category I	$285,450	$291,470	$279,840	$181,900
Category II	$269,310	$278,310	$263,470	$171,260
Category III	$264,360	$266,360	$241,890	$157,230
Category IV	$235,830	$245,200	$228,940	$148,810

Category I denotes a student in the top 25% of his class, influenced by his family, interested in sports, whose father's income is $15,000 or more.

Category II denotes a student in the top 25% of his class, influenced by his family, interested in sports, whose father's income is $10,000 to $14,999.

Category III denotes a student in the second quarter of his class, influenced by his family, interested in sports, whose father's income is $5,000 to $9,999.

Category IV denotes a student in the third quarter of his class, influenced by his family, interested in sports, whose father's income is $3,000 to $4,999.

Table Notes: The data for blacks appears two ways: I assuming no income differential and II assuming a non-white/white differential of 65%. We have assumed that all students unsure of their plans do not continue beyond high school.

The income estimates assume a productivity increase of 3% per year and a 5% discount rate. No allowance is made for price changes and/or incomes received after age 64. Although we do not include the direct costs of college attendance, the estimate of mean income includes all income recipients. Since most college students have incomes roughly equal to 25% of those not in college and since years of college attendance are included in the income figures, our estimates implicitly allow for opportunity costs foregone.[31]

Unfortunately, the 1968 census data could not be disaggregated to provide an E_{ij} for each characteristic used in the probability model so that mean earnings for each educational level (E_j) were substituted in the calculation. This did not seem to be a reasonable procedure when dealing with black incomes so we have attempted to adjust black incomes downward to reflect prevailing discrimination differentials. A reasonable estimate of the white-black income ratio is 65% based upon prior studies and we have used this estimate in column 4 of Table 4.[32]

Several other assumptions should also be noted. The Census Bureau does not provide separate estimates for graduates of business and vocational schools. Nor does it report on the income of junior college students. Income estimates for both groups are based upon the 1-3 years of college census grouping. A second problem involved our own data. The probability estimates in Table 3 include the responses of students unsure of their future plans. In calculating expected incomes a decision had to be made with respect to the estimated income of this group. We have assumed that students unsure of their plans do not continue their education. While this creates a downward bias in the income estimates for lower income groups it

should also be noted that our estimates of expected incomes of upper income groups are also biased downward since no allowance is made for different incomes from post-college training.

In general, our estimates suggest that high school graduates from upper income homes have higher expected incomes than those from lower income homes. On the average, the expected income of a student in Category IV tends to be about 83% of the income of a student from Category I and this differential would be even greater if incomes were adjusted for returns to ability.

Our estimates for the ethnic groups seem reasonable. Blacks have lower expected incomes (by about 2%) than whites even before adjusting for non-white income differentials but receive significantly lower incomes after adjustment for discrimination. Although there has been some lessening of discrimination in recent years, the non-white/white ratio may start to grow again unless the economy moves in the direction of full employment. The Cuban income estimates seem high at first glance although they are reasonable if one realizes that the Cubans in the Miami area are in the middle class and place great value on education.

Conclusion

OLS estimation of conditional probabilities of post-high school choice has several advantages. The technique is tried and true and its properties have been studied and documented. The researcher can use easily identifiable variables and, after suitable adjustment, determine probabilities directly from the regression coefficients. Moreover, the need for an

a priori specification of the model forces the researcher to give some thought to the appropriate functional form of the model.

Unfortunately, the OLS approach breaks down as the number of independent variables entered into the model increases since this inevitably leads to the multicollinearity problem. But as Glauber and Farrar point out "(s)uccessful forecasts with multicollinear variables require not only the perpetuation of a stable dependency relationship between y and X (a matrix), but also the perpetuation of stable interdependency relationships within X."[33] Since these conditions are met only in a context where the forecasting problem is trivial, this leaves several alternatives. The prediction model can be scaled down by discarding some of the prior theoretical information brought to the problem. This may involve a cost in terms of the predictive accuracy of the model. Alternatively, the researcher may reduce his data into a set of significant orthogonal common factors and use these for prediction. Where these factors can be directly identified with meaningful characteristics, this technique can provide an alternative to the one considered above. More often each factor turns out to be an artificial combination of the original variables which is difficult to identify.[34] Thus, while the use of factor analysis regression avoids the charge that too much importance is given to the effects of a single variable, it substitutes an artificial variable which may be of limited usefulness for practical applications.

In general the results obtained above using OLS techniques are encouraging. While the R^2 in our models are low, the signs on the coefficients meet our expectations and suggest the importance of including

indicators of college benefits in predictive models. Although the performance of several price variables tested in the model is disappointing, this does not rule out the possibility that once the benefits desired by students are more precisely captured in the model, a price variable will enter with a negative sign. These results suggest the need for further study of post high school choice; especially of those students choosing to continue their education but not at four-year schools.

FOOTNOTES

1. H. A. Simon, "Causal Ordering and Identifiability" in Studies in Econometric Method, ed. by W. Mood and T. Koopmans (New Haven: Yale University Press. Cowles Monograph 14), p. 50.

2. We use the term "student" to refer to all of those involved in the college decision. In fact, parents often play a major role in determining a student's post high school choice and this fact will be introduced into our later model.

3. For example, the Census Bureau frequently uses the undiscounted dollar value of a college education whereas economists prefer to speak of the present value of future earnings. For a simple explanation of the difference in the two concepts, see D. Witmer, "Economic Benefits of College Education" in Review of Educational Research, October, 1970.

4. B. Weisbrod, External Benefits of Public Education (New Jersey: Princeton University Press, 1964), p. 19. I am grateful to Weisbrod for suggesting several of these points.

5. A. Astin, The College Environment (Washington: American Council on Education, 1967).

6. J. L. Holland, "Explorations of a Theory of Vocational Choice and Achievement II. A Four Year Prediction of First Year College Performance of High Aptitude Students," Psychological Monographs, No. 570. (Washington: American Psychological Association).

7. See C. M. Grigg, W. S. Ford, H. Tuckman, D. Muse, The Demand for a Second Two Year University, A Report to the Florida International University, December, 1970.

8. H. Tuckman, High School Inputs and Their Contribution to School Performance (mimeo), Spring, 1970.

9. Grigg, et al., Op. Cit., Chap. 2.

10. See N. Draper and H. Smith, Applied Regression Analysis (New York: John Wiley, 1966), pp. 67-69.

11. John Conlisk, "Determinants of School Enrollment and School Performance," Journal of Human Resources, Spring, 1969.

12. See, for example, B. Bloom and F. Peters, The Use of Academic Prediction Scales for Counseling and Selecting College Entrants (New York: Free Press of Glencoe, 1961).

13. Grigg, Op. Cit., Chap. 4.

14. See Appendix Table 3.

15. Grigg, Op. Cit., Chap. 2.

16. See, for example, Grigg, Chap. 4.

17. S. Hoenack, Private Demand for Higher Education in California (unpublished dissertation, University of California, Berkeley, 1969).

18. See S. Hoenack, Op. Cit., Chap. 3, 4; W. Sewell and V. Shah, "Socioeconomic Status, Intelligence, and the Attainment of Higher Education," Sociology of Education, Fall 1967; and R. Radner and S. Miller, "Demand and Supply In U. S. Higher Education: A Progress Report," American Economic Review, May 1970.

19. The use of a dichotomous variable as the dependent variable violates the assumption of homoscedasticity of the OLS model. This means our estimates are not as efficient as those obtained when all assumptions are met. Given time constraints and other limitations in the model, the additional effort required for more refined techniques was not thought to be justified.

20. Beta coefficients can be computed from the formula $B_i^* = b_i \frac{sx_i}{s_y}$ where sx_i is the standard deviation of the x_i parameter and s_y represents the standard deviation for the dependent variable.

21. Guy Orcutt, Microanalysis of Socioeconomic Systems: A Simulation Study (New York: Harper, 1961).

22. More formally, for each category we have $\frac{1}{n} \sum_{i=1}^{n} (y_i - y_i^*) = \bar{e}$. Unfortunately, the choice of these groupings is arbitrary since Orcutt's only restriction is that they be homogeneous. As a result, the conditional probabilities obtained by the researcher are not invariate to the residual groups chosen.

23. See H. Tuckman, Op. Cit.

24. Grigg, et al, Op. Cit., especially Chap. 5.

25. L. Schoenfeldt, "Education After High School," Sociology of Education, Fall 1968.

26. Ibid., p. 359.

27. Examples of this approach are found in Conlisk, Op. Cit., Tuckman, Op. Cit., and the model presented above.

28. See, for example, W. Hansen, B. Weisbrod, and W. S. Scanlon, "Schooling and Earnings of Low Achievers," American Economics Review, June, 1970 and J. Gwartney, "Changes in the Non-White/White Income Ratio 1939-67," American Economic Review, December, 1970.

29. U. S. Department of Commerce, Bureau of the Census, "Annual Mean Income, Lifetime Income, and Educational Attainment of Men in the United States For Selected Years, 1956 to 1968," Current Population Reports, Series P-60, No. 74, October 30, 1970.

30. The formula used to calculate the estimates is

$$V_{18} = \frac{\sum_{N=18}^{64} Y_n P_n (1+x)^{N-18+1/2}}{(1+R)^{N-18+1}}$$

Where V stands for the presented value of income received from age 18 to 64, Y_n indicates average income at age N, P_n gives the relative number of survivors at age N of those alive at age 18 (based on U. S. life tables), x represents a 3% productivity increase, and R stands for a 5% discount rate. For further information, see "Annual Mean Income...," pp. 18-19.

31. The present value of a college graduate's income is $297,000 at age 18 and $300,000 at age 22. The difference between the two estimates reflects the effects of beginning the discounting at a time when incomes are low and applying larger discount rates in periods when incomes rise.

32. Gwartney, Op. Cit., p. 874 and D. O'Neill, "The Effect of Discrimination on Earnings: Evidence from Military Test Score Results," Journal of Human Resources, Fall, 1970, especially p. 482-484. Black incomes tend to be less than those of whites in later years but since these years are heavily discounted this difference should not affect our estimates.

33. D. Farrar & R. Glauber, "Multicollinearity In Regression Analysis: The Problem Revisited," Review of Economics and Statistics, February, 1967.

34. Ibid, p. 97.

APPENDIX TABLE 1

Means and Standard Deviations of Variables Used In the Educational Choice Models

Variable Name		Mean	Standard Deviation
Male	(x_1)	.499	.500
Black	(x_2)	.153	.360
Cuban	(x_3)	.112	.315
Class Rank			
Top 25% of Class	(x_4)	.266	.442
25-50% of Class	(x_5)	.256	.436
50-75% of Class	(x_6)	.319	.466
Education of Father			
Some College Education	(x_7)	.208	.406
College Graduate	(x_8)	.206	.405
Family Income			
$3,000-4,999	(x_9)	.084	.277
$5,000-9,999	(x_{10})	.216	.411
$10,000-14,999	(x_{11})	.247	.432
Above $15,000	(x_{12})	.259	.438
Activity Desired by Student			
Social	(x_{13})	.128	.335
Special Interest	(x_{14})	.088	.283
Political Activity	(x_{15})	.027	.161
Future Career	(x_{16})	.101	.301
Sports	(x_{17})	.172	.378
Other College Related	(x_{18})	.049	.216
Major Influence on the Student			
Father or Mother	(x_{19})	.612	.487
Counselor or Teacher	(x_{20})	.057	.233
Friend	(x_{21})	.105	.307
Relative or Other	(x_{22})	.092	.290

Original Set of Variables

1. Sex
2. Class
3. Color
4. Top Quarter of Class
5. Second Quarter of Class
6. Third Quarter of Class
7. Father has some College Education
8. Father is College Graduate
9. Father's Income is $1,000-4,999
10. Father's Income is $5,000-9,999
11. Father's Income is $10,000-14,999
12. Father's Income is $15,000+
13. Social Science Major
14. Fine Arts Major
15. Science Major
16. Education Major
17. Business Major
18. Other Major
19. Social Activities
20. Special Interest Activities
21. Political Activities
22. Future Career Activities
23. Sports Activities
24. Other Activities
25. Father or Mother is Major Influence
26. Counselor or Teacher is Major Influence
27. Friend is Major Influence
28. Relative or Other is Major Influence
29. Value of Education is Financial
30. Value of Education is Knowledge
31. Value of Education is Civic
32. Value of Education is Other
33. Tuition of First College Considered
34. Tuition Plus Fees of First College Considered
35. Tuition of Second College Considered
36. Tuition Plus Fees of Second College Considered
37. Tuition times x_9
38. Tuition times x_{10}
39. Tuition times x_{11}
40. Tuition times x_{12}

APPENDIX TABLE 3

Matrix of Zero Order Correlations

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	x_{21}	x_{22}
x_1	1.00	-.03	-.00	-.02	.04	-.01	.00	.03	.01	-.03	.10	.04	-.07	-.07	.00	.00	.23	-.01	.03	-.04	-.08	-.04
x_2		1.00	-.15	-.12	.01	.01	-.15	-.16	.15	.02	-.01	-.17	-.01	.01	-.02	.00	.01	-.07	.04	.05	-.05	.00
x_3			1.00	-.07	.02	.05	-.01	-.01	.14	.17	-.08	-.13	-.03	.02	-.00	.00	.01	-.01	.05	.01	-.03	-.05
x_4				1.00	-.35	-.41	.08	.12	-.06	-.02	.04	.07	.09	.08	.08	.02	.05	.05	.08	.02	-.05	-.00
x_5					1.00	-.40	.00	.00	.01	.01	.01	.00	.02	-.02	.04	.00	.03	.02	.01	.02	.01	-.01
x_6						1.00	.00	-.08	.02	.02	.02	-.03	.01	-.06	-.05	-.01	-.03	-.03	-.02	-.05	.02	.04
x_7							1.00	-.26	-.05	-.04	.10	.07	.03	.02	-.00	.05	.01	.01	.05	-.01	.00	-.03
x_8								1.00	-.10	-.08	-.07	.25	.04	.01	.07	-.00	.06	-.01	.09	-.07	-.00	-.00
x_9									1.00	-.16	-.17	-.18	-.05	-.00	-.02	.01	.01	-.06	.01	.02	.03	.04
x_{10}										1.00	-.30	-.31	.02	.01	-.03	-.03	.03	.01	-.01	.08	.02	-.00
x_{11}											1.00	-.34	.00	-.01	-.00	.02	.04	.03	.02	-.01	.00	.02
x_{12}												1.00	.07	.03	.04	-.00	-.01	.02	.06	-.06	.01	-.03
x_{13}													1.00	-.12	-.06	-.13	-.18	-.09	.06	-.00	-.01	-.01
x_{14}														1.00	-.05	-.10	-.14	-.07	.03	.02	.02	-.03
x_{15}															1.00	-.06	-.08	-.04	.01	-.01	-.01	.04
x_{16}																1.00	-.15	-.08	.03	.01	-.06	-.04
x_{17}																	1.00	-.10	.12	.01	-.05	-.04
x_{18}																		1.00	.01	.02	-.01	.01
x_{19}																			1.00	-.31	-.43	-.40	
x_{20}																				1.00	-.02	-.06
x_{21}																					1.00	-.11